Computers and Artificial Intelligence, Vol. 17, 1998, No. 1, 1—

INFORMATION RETRIEVAL, IMAGING AND PROBABILISTIC LOGIC

Fabrizio SEBASTIANI

Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 – 56126 Pisa (Italy)
E-mail: fabrizio@iei.pi.cnr.it

Abstract. Imaging is a class of non-Bayesian methods for the revision of probability density functions originally proposed as a semantics for conditional logic. Two of these revision functions, standard imaging and general imaging, have successfully been applied to modelling information retrieval by Crestani and van Rijsbergen. Due to the problematic nature of a"direct" implementation of imaging revision functions, in this paper we propose their alternative implementation by representing the semantic structure that underlies imaging-based conditional logics in the language of a probabilistic (Bayesian) logic. Besides showing the potential of this "Bayesian" tool for the representation of non-Bayesian revision functions, recasting these models of information retrieval in such a general purpose knowledge representation and reasoning tool paves the way to a possible integration of these models with other more KR-oriented models of IR, and to the exploitation of general-purpose domain-knowledge.

Keywords: probability kinematics, probabilistic reasoning, information retrieval, imaging

1 INTRODUCTION

In recent years researchers have devoted an increasing effort to the specification of models of information retrieval (IR) along the so-called *logical approach* [20]. Although there exist various interpretations of this approach, by and large we may take it to say that the relevance of documents to user queries may be viewed in terms of the validity of the formula $d \rightarrow q$ of a logical language, where d is a formula representing the document, q a formula representing the query and "$\rightarrow$"

is the conditional ("implication") connective of the logic in question[1]. This way of viewing IR is especially fascinating once we consider, instead of the proof-theoretic, "symbol-crunching" level of logic, its model-theoretic, semantic level. In terms of the latter, the logical approach to IR amounts to sanctioning that relevance coincides with (set-)inclusion of information content, or semantics: only documents whose information content includes that of the query are to be retrieved.

However, the practical impossibility of finding *perfect* (i.e. absolutely faithful) representations of the information content of documents and queries calls for a probabilistic treatment of this conditional sentence: it is then generally acknowledged that a realistic approach to the IR problem must rather rely on the evaluation of the real-valued term $P(d \rightarrow q)$, where $P(\alpha)$ stands for "the probability that α". According to this position, then, the definition of a model of IR involves:

1. the definition of a model for documents, i.e. of a methodology that, given a document **d**, produces a logical formula d that constitutes a symbolic representation of it;
2. analogously, the definition of a model for queries;
3. the choice of a suitable base (i.e. non-probabilistic) logic, i.e. one in which the formula $d \rightarrow q$ is valid if and only if the document represented by d is relevant to the query represented by q *under the idealised assumption that* d *and* q *are perfect representations of the document and the query, respectively*;
4. the choice of an approach to the representation of probability, i.e. of a way of relaxing the idealised assumption described above and producing a ranking of documents according to the probability of their relevance. Accordingly, one will require that $P(d_1 \rightarrow q) \geq P(d_2 \rightarrow q)$ holds if and only if the document represented by d_1 is more likely to be relevant to the query represented by q than the document represented by d_2.

A number of researchers have recently taken up these ideas, and proposed logics and logic-based models of IR based on them. Among these, of particular interest to the present paper are the models of IR based on "imaging" [12] (hereafter called *standard* imaging) and "general imaging" [7] by Crestani and van Rijsbergen [4, 5, 21]. Standard imaging and general imaging are *density revision functions* (DRFs – see Section 2) originally proposed as a semantics for *conditional logic*, the branch of logic that addresses the "if ... then ..." notion of natural language. The experimental results presented in [4, 5] show a definite improvement of performance over standard approaches to IR, thus supporting the conjecture that imaging methods capture some fundamental intuition underlying IR.

A full-blown implementation of imaging methods is, unfortunately, problematic. The reason is that implementation techniques for DRFs (of which belief networks

[1] For a discussion why we think that *validity*, rather than *truth*, of $d \rightarrow q$ is the notion to consider, see [7].

are a primary example) have so far concentrated on the Bayesian case, i.e. *Bayesian conditionalisation*. To our knowledge, no technique has been developed yet for non-Bayesian DRFs such as imaging, and no theorem proving technique has been developed for imaging-based conditional logics. In this paper we propose an alternative method for implementing imaging methods. Essentially, the idea is to represent the semantic structure that underlies imaging-based conditional logics in the language of a probabilistic (Bayesian!) logic. This process of *abstraction* (i.e. of transfer from the realm of semantics to that of syntax) is conceptually not dissimilar from the so-called "standard translation" (see e.g. [19]) of modal propositional logic into first order logic (FOL), whereby modal propositional reasoning is reduced to standard FOL reasoning by simulating within FOL the possible worlds semantics of modal propositional logic.

We show that Halpern's $\mathcal{L}_1$ [9] logic, a simple FOL extended with features for *objective* probability, is powerful enough to accommodate not only standard and general imaging, but also generalizations of them such as "proportional imaging" (see Section 2). In another paper [3] we also show that an extension of $\mathcal{L}_1$ with features for *subjective* probability (called $\mathcal{L}_3$) can further accommodate "Jeffrey imaging", a variant of imaging obtained by combining (any variant of) imaging and *Jeffrey conditionalisation* [11] which seems a promising tool for the analysis of non-binary "relevance feedback" in IR [2]. Our implementation of imaging (and variations thereof) on top of $\mathcal{L}_1$ shows then that Bayesian revision tools can be seen as convenient and powerful toolkits for fast prototyping of non-Bayesian models of IR. Quite obviously, recasting these models of IR in such a general purpose knowledge representation (KR) and reasoning tool paves the way to a possible integration of these models with other more KR-oriented models of IR (such as e.g. [13, 16]), and to the exploitation of general-purpose domain-knowledge.

The paper is organised as follows. While in Section 2 we briefly review imaging DRFs, in Section 3 we look at the main features of the $\mathcal{L}_1$ probabilistic logic, the main tool that we will use in this work. In Section 4 we show $\mathcal{L}_1$ implementations of models of IR based on standard imaging, general imaging and proportional imaging. Section 5 discusses both some theoretical underpinnings and the practical consequences of our work by comparing it with related work.

2 THE BAYESIAN MODEL OF EPISTEMIC STATES

The notion of imaging (together with its variations) assumes that the epistemic state of a cognitive agent is represented by a (subjective) *probability function* P defined on the set of sentences of a language L (see e.g. [8, pages 36–40]) and that complies with the standard axioms of probability. If A is a sentence of L, then $P(A)$ is meant to represent the *degree of confidence* (or certainty, or belief) that the agent has in the truth of A: if $P(A) = 1$, the agent is certain of the truth of A; if $P(A) = 0$, the agent is certain that A is false, while if $0 < P(A) < 1$ the agent

is unsure whether A is the case or not. From now on, we will take L to be the language of propositional logic defined on a finite number of propositional letters.

In real-life situations, agents may change their mind as a result of the acquisition of new evidence; e.g. the agent may come to believe true facts that she believed to be probably false. In order to model this, one needs a mechanism to change the probability value associated to a sentence and change the values of other semantically related sentences accordingly. This is called a probability *revision* function[2]. The standard probability revision function is *Bayesian conditionalisation*, according to which if an agent comes to firmly believe in the truth of a sentence A (which she believed to be at least possibly true — i.e. $P(A) > 0$), her new epistemic state must be described by a new probability function $P(-|A)$ (also indicated as P_A) which, for all sentences B, is defined as:

$$P_A(B) \stackrel{\text{def}}{=} P(B|A) \stackrel{\text{def}}{=} \frac{P(A \wedge B)}{P(A)}. \tag{1}$$

Note that the new probability function is such that A is (correctly) deemed true, i.e. $P(A|A) = 1$. An alternative, semantically oriented but equivalent way of characterizing epistemic states is to assume thatthere is a *density function* (also called a probability *distribution*) μ on the set W of the 2^n *possible worlds* (or simply *worlds*) on which the propositional language is interpreted (where n is the number of propositional letters in the language)[3]; i.e. μ is such that $\sum_{\{w \in W\}} \mu(w) = 1$. The degree of confidence of the agent in sentence A is defined as the sum of the probabilities of the worlds that satisfy A ("A-worlds"), i.e.

$$P(A) \stackrel{\text{def}}{=} \sum_{\{w \in W \ | \ w \models A\}} \mu(w) \tag{2}$$

In the possible worlds view, instead of specifying a probability revision function one specifies a *density revision function* (i.e. a function mapping a density function into another density function) on possible worlds that induces the desired probability revision function through (2). Viewed as a DRF, Bayesian conditionalisation then

[2] In the literature (e.g. [8]) a distinction is made between probability *expansion*, *revision* and *contraction* functions, depending on the probability values associated to the sentence under consideration before and after the change. In what follows we will avoid this finer-grained distinction and use the term "revision" to collectively indicate all three kinds of transformation.

[3] This characterisation of possible worlds should not be confused with the Hintikka-Kripke notion, according to which for giving semantics to modal logic we group possible worlds in so-called "Kripke structures", and each of them is considered "possible" by an explicitly (rather than implicitly) represented cognitive agent. This latter notion implies a notion of belief located in the object language (with explicit "$\mathbf{Bel}_i$" operators, where "$\mathbf{Bel}_i(\alpha)$" means "agent i believes that α"), rather than in the metalanguage as the view we discuss.

amounts to eliminating from consideration the worlds that do not satisfy A ("$\neg A$-worlds"), and creating a new density function μ' obtained from μ by redistributing to the A-worlds the probability originally assigned to the $\neg A$-worlds, where the redistribution is proportional to the probability originally assigned to the A-worlds. Therefore, the revised density function μ' will be such that:

$$\mu'(w) = \begin{cases} \mu(w) \cdot (1 + \dfrac{P(\neg A)}{P(A)}) & \text{if } w \models A \\ 0 & \text{if } w \not\models A. \end{cases} \tag{3}$$

Imaging and its variations are DRFs alternative to Bayesian conditionalisation. They differ from it in that they are based on the idea that the probability of $\neg A$-worlds is *not* redistributed proportionally to the original probability of the A-worlds. The underlying assumption is that there is a measure S of *similarity* defined on W such that $0 \leq S(w, w') \leq 1$ measures, for every pair $\langle w, w' \rangle \in W^2$, how is w' similar to w[4]. According to imaging DRFs, only worlds sufficiently similar to the $\neg A$-worlds receive some probability; exactly, the fact how similar they need to be in order to receive some probability is what differentiates the various forms of imaging.

Standard imaging (first introduced in [12]) is based on the simplifying assumption that, for all satisfiable sentences A and for all $\neg A$-worlds w, a most similar A-world $w' = \sigma(A, w) \stackrel{\text{def}}{=} \max\{S(w, w') \mid w' \models A\}$ always exists and is unique[5]; it is to w' that the probability $\mu(w)$ is transferred. Imaging thus sanctions that:

$$\mu'(w') = \begin{cases} 0 & \text{if } w' \not\models A \\ \mu(w') + \displaystyle\sum_{\{w \in W \mid w' = \sigma(A,w)\}} \mu(w) & \text{if } w' \models A \end{cases} \tag{4}$$

Obviously, the results of applying imaging depend on the choice of the S function. In the general case, however, for no choice of the similarity function the results of Bayesian conditionalisation coincide with those of imaging.

A generalisation of this method, called *general imaging* (first introduced in [7] – see also [8, pages 108–117]), is based on the idea of relaxing the uniqueness

[4] The higher the value of $S(w, w')$, the more is w similar to w'. Whether similarity needs to be a symmetric measure is not discussed by the proponents of Imaging methods; we will thus refer to the general case in which it is not.

[5] In the discussion of imaging and its variations we will always assume that any possible world w that receives some probability in the transfer process is such that $\mu(w) > 0$; therefore, in the following definitions the clause "if $w' \models A$" should actually be read as "if $w' \models A$ and $\mu(w) > 0$", and the clause "if $w' \not\models A$" should actually be read as "if $w' \not\models A$ or $\mu(w) = 0$". This requirement is necessary for imaging methods to be "preservative", a common requirement for probability revision functions. As argued in [8, page 118], "the only interesting form of imaging is preservative imaging".

assumption for the most similar A-world. For each satisfiable sentence A and for each $\neg A$-world w, we now have a *set* of (equally) most similar A-worlds $W = \sigma(A, w)$, among which the probability $\mu(w)$ is thus distributed on an equal basis. General imaging thus sanctions that:

$$\mu'(w') = \begin{cases} 0 & \text{if } w' \not\models A \\ \mu(w') + \sum\limits_{\{w \in W \mid w' \in \sigma(A,w)\}} \frac{\mu(w)}{|\sigma(A,w)|} & \text{if } w' \models A \end{cases} \tag{5}$$

where $|\sigma(A, w)|$ indicates the cardinality of the set $\sigma(A, w)$.

In standard and general imaging we do *not* need to assume that there is a quantitative measure of similarity between possible worlds. What we only need to assume (qualitatively) is that, given world w, in the set of A-worlds there is a most similar world (standard imaging) or a distinguished subset of most similar worlds (general imaging).

The quantitative assumption is instead necessary in a further generalisation of this method, which we may call for convenience *proportional imaging*. This can be obtained by assuming that the probability of $\neg A$-worlds is distributed not among a (usually small) set of most similar A-worlds, but among the set of *all* A-worlds, in a way that is directly proportional to the degree of similarity between the "donor" and the "recipient". Proportional imaging thus sanctions that:

$$\mu'(w') = \begin{cases} 0 & \text{if } w' \not\models A \\ \mu(w') + \sum\limits_{\{w \in W \mid w \neg\models A\}} \mu(w) \cdot \frac{S(w,w')}{\sum\limits_{\{w'' \in W \mid w'' \models A\}} S(w,w'')} & \text{if } w' \models A \end{cases} \tag{6}$$

Combinations between general and proportional imaging may be defined by sanctioning that probability be distributed in the style of proportional imaging, but to a subset of the A-worlds only. For instance, we may decide probability to be distributed only to those A-worlds whose similarity to the donor is higher than a threshold value k (when such worlds exist; otherwise, the most similar worlds are selected independently of their similarity to the donor). Therefore, if we take the set $\sigma(A, w)$ to be

$$\sigma(A,w) = \begin{cases} \max\{S(w,w') \mid w' \models A\} & \text{if } |\{w' \in W | w' \models A \wedge \\ & \quad \wedge S(w,w') \geq k\}| = 0 \\ \{w' \in W | w' \models A \wedge S(w,w') \geq k\} & \text{otherwise} \end{cases} \tag{7}$$

we have

$$\mu'(w') = \begin{cases} 0 & \text{if } w' \not\models A \\ \mu(w') + \sum\limits_{\substack{\{w \in W \mid w \not\models A \ \wedge \\ \wedge \ w' \in \sigma(A,w)\}}} \mu(w) \cdot \frac{S(w,w')}{\sum\limits_{w'' \in \sigma(A,w)} S(w,w'')} & \text{if } w' \models A \end{cases} \tag{8}$$

Alternatively, we may decide probability to be distributed in the style of proportional imaging, but only to the s A-worlds most similar to the donor (if s A-worlds exist at all; otherwise the probability is redistributed to all the A-worlds). In this case, the set $\sigma(A, w)$ becomes

$$\sigma(A, w) = \begin{cases} \max_s\{S(w, w') \mid w' \models A\} & \text{if } |\{w' \in W \mid w' \models A\}| > s \\ \{w' \in W \mid w' \models A\} & \text{otherwise} \end{cases} \quad (9)$$

where $\max_s\{S(w, w') \mid w' \models A\}$ denotes the set of the s A-worlds most similar to w, and $\mu'(w')$ is again computed according to (8).

3 THE $\mathcal{L}_1$ PROBABILISTIC LOGIC

The $\mathcal{L}_1$ probabilistic logic is a first order logic for reasoning about (objective) probabilities [9]. Probability values can explicitly be mentioned in the language: rather than mapping non-probabilistic formulae on the real interval $[0, 1]$, probabilistic formulae are mapped on the standard truth values *true* and *false*. The logic allows the expression of real-valued terms of type $w_{\langle x_1,\ldots,x_n\rangle}(\alpha)$ (where α is any $\mathcal{L}_1$ formula), with the meaning "the probability that random individuals $x_1, \ldots, x_n$ verify α". It also allows their comparison by means of standard numerical binary operators, resulting in formulae that can be composed by the standard sentential operators of first order logic. The semantics of the logic is given by assuming the existence of a discrete probability structure on the domain; a formula such as $w_{\langle x_1,\ldots,x_n\rangle}(\alpha) \geq r$ is true in an interpretation iff the probability assigned to the individuals that verify α sums up to at least r[6].

The semantics of $\mathcal{L}_1$ can be specified by means of *type 1 probabilistic structures* (PS_1), i.e. triples $M = \langle D, \pi, \mu\rangle$, where D is a domain of individuals, π is an assignment of n-ary relations on D to n-ary predicate symbols and of n-ary functions on D to n-ary function symbols ($\langle D, \pi\rangle$ is then a first order interpretation), and μ is a discrete density function (DPD) on D.

The numerical value $\mu(d)$ may be interpreted as "the probability that, if a random individual has been picked from the domain D, it is d". In what follows, we will use $\mu(D')$ (where $D' \subseteq D$) as a shorthand for $\sum_{d\in D'} \mu(d)$. Also, given a DPD μ on D, μ^n is defined as the DPD on D^n such that $\mu^n(\langle d_1, \ldots, d_n\rangle) = \mu(d_1) \times \ldots \times \mu(d_n)$.

A *valuation* is a mapping v of object variables (i.e. variables denoting individuals of the domain, indicated by the superscript ^{o}) into D and numerical variables (i.e. variables denoting real numbers, indicated by the superscript ^{c}) into R. Three semantic notions can now be defined:

[6] It follows that, if x does not occur free in α, the term $w_{\langle x\rangle}(\alpha)$ may evaluate to 0 or 1 only, depending whether α evaluates to *false* or *true*, respectively. Given a closed formula α, the term $w_{\langle x\rangle}(\alpha)$ plays then the role of its characteristic function.

- the *(numerical) value* $\|t^c\|_{\langle M,v\rangle}$ of a numerical term t^c in $\langle M, v\rangle$, with values in the real interval $[0, 1]$;
- the *(object) value* $[t^o]_{\langle M,v\rangle}$ of an object term t^o in $\langle M, v\rangle$, with values in D;
- the *truth* $\langle M, v\rangle \models \alpha$ of a formula α in $\langle M, v\rangle$, with values in $\{true, false\}$.

The semantics of the logic is more formally described by the semantic clauses that follow. In these, "*mathop*" is an operator in the set $MATHOP$=$\{+,-,\cdot,\div\}$, and "*relop*" is an operator in the set $RELOP$=$\{=,\neq,\geq,\leq,<,>\}$; **mathop** and **relop** are the corresponding operations on real numbers.

$$
\begin{aligned}
[x^o]_{\langle M,v\rangle} &= v(x^o)\\
[f_i^n(t_1^o,\ldots,t_n^o)]_{\langle M,v\rangle} &= \pi(f_i^n)([t_1^o]_{\langle M,v\rangle},\ldots,[t_n^o]_{\langle M,v\rangle})\\
\|x^c\|_{\langle M,v\rangle} &= v(x^c)\\
\|k^c\|_{\langle M,v\rangle} &= \mathbf{k}\\
\|t_1^c \ mathop\ t_2^c\|_{\langle M,v\rangle} &= \|t_1^c\|_{\langle M,v\rangle}\ \mathbf{mathop}\ \|t_2^c\|_{\langle M,v\rangle}\\
\|w_{\langle x_1^o,\ldots,x_n^o\rangle}(\alpha)\|_{\langle M,v\rangle} &= \mu^n(\{\langle d_1,\ldots,d_n\rangle | \langle M, v[x_1^o/d_1,\ldots,x_n^o/d_n]\rangle \models \alpha\})\\
\langle M, v\rangle \models P_i^n(t_1^o,\ldots,t_n^o) &\text{ iff } \langle [t_1^o]_{\langle M,v\rangle},\ldots,[t_n^o]_{\langle M,v\rangle}\rangle \in \pi(P_i^n)\\
\langle M, v\rangle \models \neg\alpha &\text{ iff } \langle M, v\rangle \not\models \alpha\\
\langle M, v\rangle \models \alpha \wedge \beta &\text{ iff } \langle M, v\rangle \models \alpha \text{ and } \langle M, v\rangle \models \beta\\
\langle M, v\rangle \models \forall x^o.\alpha &\text{ iff } \langle M, v[x^o/d]\rangle \models \alpha \text{ for all } d \in D\\
\langle M, v\rangle \models \forall x^c.\alpha &\text{ iff } \langle M, v[x^c/r]\rangle \models \alpha \text{ for all } r \in \mathcal{R}\\
\langle M, v\rangle \models t_1^c \ relop\ t_2^c &\text{ iff } \|t_1^c\|_{\langle M,v\rangle}\ \mathbf{relop}\ \|t_2^c\|_{\langle M,v\rangle}\\
\langle M, v\rangle \models t_1^o = t_2^o &\text{ iff } [t_1^o]_{\langle M,v\rangle}=[t_2^o]_{\langle M,v\rangle}
\end{aligned}
$$

A formula α is *satisfiable* iff there exists $\langle M, v\rangle$ such that $\langle M, v\rangle \models \alpha$; a formula α is *valid* (in symbols: $\models \alpha$) iff $\langle M, v\rangle \models \alpha$ for all $\langle M, v\rangle$. Validity, the main notion of interest in reasoning contexts, has been shown to be decidable in $\mathcal{L}_1$ when the domain D has a fixed, finite cardinality n (see [9]). Note that, although the syntax of the logic might seem too limited for practical uses, a number of other constructs may be defined as "shorthands" of the above formulae. For instance, the Bayesian conditionalisation operator "$w_{\langle x_1,\ldots,x_n\rangle}(-|-)$" is expressed by considering the formula $w_{\langle x_1,\ldots,x_n\rangle}(\alpha|\beta) = r$ as shorthand for the formula $w_{\langle x_1,\ldots,x_n\rangle}(\alpha \wedge \beta) = r \cdot w_{\langle x_1,\ldots,x_n\rangle}(\beta)$. Similarly, the square root operator "$\sqrt{-}$" is expressed by considering the formula $\sqrt{t^c} = r$ as shorthand for the formula $t^c = r \cdot r$. In an actual implementation of the logic, numerical functions such as "$\sqrt{-}$" can obviously be implemented as calls to appropriate subroutines rather than as expansions into the appropriate axiomatic definitions, which then serve for theoretical purposes only.

4 A REPRESENTATION OF IMAGING ON TOP OF PROBABILISTIC LOGIC

Crestani and van Rijsbergen's models of IR are based on a somewhat non-standard interpretation of imaging DRFs, as[7]

1. he representation language is not that of propositional logic but a language of simple propositional letters, each representing a document or a query;
2. ossible worlds are keywords; this means that there are not necessarily 2^n possible worlds, but there are as many possible worlds as the number of keywords in the application domain. The propositional letter d_i (resp. q_i) is conventionally taken to be true at world t_j iff the document represented by d_i (resp. the query represented by q_i) is indexed by the keyword represented by t_j.

We now describe a representation of the models of IR of [4, 5] in terms of $\mathcal{L}_1$. Our purpose is to show how the representation of these mechanisms may be accomplished quite easily, thus establishing Bayesian tools as convenient and powerful platforms for fast prototyping of non-Bayesian IR models. To this end, we do not confine ourselves to the characterisation of just the models presented in [4, 5], but go on to show how some generalisations of them can also be easily represented.

In this approach, the whole information retrieval process is modelled as *a proper theory* of $\mathcal{L}_1$, obtained by assembling together various sets of formulae, each representing a class of entities participating in the process. In Section 4.1 we describe in detail the characterisation of the IR model based on (standard) imaging, while in Sections 4.2 and 4.3 we describe the modifications we need to make to it in order to transform it into a characterisation of an IR model based on general imaging and proportional imaging, respectively.

4.1 Representing (Standard) Imaging

In order to implement standard imaging, a first subset of $\mathcal{L}_1$ formulae is necessary to identify keywords and documents. This is necessary, as the domain of interpretation must be restricted to deal with these types of individuals only, which are the only entities of interest in the revision processes. Assuming that $\{t_1, \ldots, t_n\}$ is the language of keywords by means of which documents are represented, and that $\{d_1, \ldots, d_m\}$ are the documents in our collection, we need the formulae

$$Keyword(t_1) \ \wedge \ldots \wedge \ Keyword(t_n) \tag{10}$$

$$Document(d_1) \ \wedge \ldots \wedge \ Document(d_m) \tag{11}$$

$$\forall x.[x = t_1 \vee \ldots \vee x = t_n \vee x = d_1 \vee \ldots \vee x = d_m] \tag{12}$$

$$\forall x.\neg(Document(x) \wedge Keyword(x)) \tag{13}$$

[7] The actual methods with which Crestani and van Rijsbergen have dealt with are (standard) imaging (in [4]) and an approximation of the combination of general and proportional imaging described by Equation 9 (in [5]).

This is a key feature of this approach: documents and keywords are individuals *belonging to* the domain of discourse of a first order interpretation, while in [4, 5]'s original approach keywords *are* (propositional) interpretations and documents are propositions. Back to this point in Section 5.

The next subset of formulae is the one that specifies keyword occurrence, i.e. which documents are indexed by which keywords. We represent this by formulae

$$w_x(Occ(t_i, d_j)) = o_{ij} \qquad o_{ij} \in \{0, 1\} \tag{14}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where o_{ij} is 1 iff t_i occurs in d_j. This representation is made possible by the fact that, as noted in Footnote 6, the probability operator applied to a closed formula yields the formula's characteristic function.

Next, the probability of each keyword t_i is specified by means of the set of formulae

$$w_x(x = t_i \mid Keyword(x)) = p_{t_i} \qquad p_{t_i} \in [0, 1] \tag{15}$$

for all $i = 1, \ldots, n$. These formulae account for the case in which we want to input the probability values p_{t_i} from the outside. Alternatively, these probability values can be computed *within* $\mathcal{L}_1$ from the already available occurrence data, e.g. as their inverse document frequency (IDF – see e.g. [15]). In this case, the formulae (15) are substituted by the formulae

$$w_x(x = t_i \mid Keyword(x)) = -log(w_y(Occ(t_i, y) \mid Document(y))) \tag{16}$$

Formulae (16) compute the probabilities of keywords as their inverse document frequency; in fact, the formula $w_y(Occ(t_i, y) \mid Document(y))$ is to be read as "the probability that, by picking a random document y, keyword t_i occurs in y". For (16) to truly represent IDF, though, we must assume that documents are picked with equal probability, which we state by the formula

$$\forall xy.(Document(x) \wedge Document(y)) \Rightarrow [w_z(x = z) = w_z(y = z)] \tag{17}$$

Alternatively, one might choose to include both formulae (15), (16) and (17) in the representation. In this way, probability values would be precomputed "externally" and input to the reasoning process through formulae (15), and formulae (16) and (17) would act as *integrity constraints*. In what follows we will use the expression $P(t_i)$ as a shorthand of the expression $w_x(x = t_i \mid Keyword(x))$.

The next subset of formulae specifies the *similarity matrix*, i.e. how similar document d_i is to document d_j for all $1 \leq i, j \leq m$, $i \neq j$:

$$Sim(t_i, t_j) = s_{ij} \qquad 1 \leq i, j \leq m, i \neq j \tag{18}$$

Only similarities between nonequal documents are specified; in fact, the case $i = j$ is not interesting for imaging methods, and its specification would complicate

the expression of formulae (23). Values s_{ij} are input from an external source of information. Alternatively, they can be computed from within $\mathcal{L}_1$ from the already available occurrence values; for instance, they may be taken to be equivalent to the degree of coextensionality of the Occ predicate and computed by means of the formula

$$Sim(t_i, t_j) = w_x(Occ(t_i, x) \mid Occ(t_j, x)) \cdot w_x(Occ(t_j, x) \mid Occ(t_i, x)) \tag{19}$$

or else be computed according to some other measure of similarity (e.g. the EMIM measure adopted in [4]). Again, formulae (18) and (19) might coexist, with formulae (19) acting then as integrity constraints. Further integrity constraints might be added, if one's theory of similarity requires one to do so, in order to state further properties of similarity; e.g. similarity may be constrained to be a symmetric relation:

$$\forall xy.[Sim(x, y) = Sim(y, x)] \tag{20}$$

and/or a triangular relation:

$$\forall xy.[Sim(x, y) + Sim(y, z) \geq Sim(x, z)] \tag{21}$$

The following subset of formulae specifies, for each keyword, its most similar keyword:

$$MostSim(t_i, t_{k_i}) \qquad 1 \leq i \leq n. \tag{22}$$

Similarly to formulae (15) and (18) these formulae account for the case in which we want to input the "most-similarity" values from outside. Alternatively, these values can be computed within $\mathcal{L}_1$ from the already available similarity data by means of the formulae

$$MostSim(t_i, t_{k_i}) \Leftrightarrow \neg\exists t_j.[Sim(t_i, t_j) \geq Sim(t_i, t_{k_i})]. \tag{23}$$

Again, formulae (22) and (23) may coexist, with formulae (23) acting then as integrity constraints.

Next, we have to show how to calculate the revised probability of keyword t_i by imaging on document d_j, i.e. how to implement the probability transfer function. The revised probabilities are specified by the following numerical terms, for $1 \leq i \leq n$:

$$\begin{aligned} & w_x(Occ(t_i, d_j)) \cdot [P(t_i) + \\ & \quad + \sum_{k=1}^{n} [P(t_k) \cdot w_x(\neg Occ(t_k, d_j)) \cdot w_x(MostSim(t_k, t_i))] \end{aligned} \tag{24}$$

To interpret term (24) remember Footnote 6 and note that all formulae occurring in the context of a w_x operator are closed (w_x-terms thus act here as "guards"). The summation operator $\sum$ is obviously a shorthand for the corresponding expanded

numerical term. In what follows we will use the expression $P^{\#}_{d_j}(t_i)$ as a shorthand of expression (24).

In order to compute relevance of documents to the query, we now have to indicate by which keywords the query q is indexed. This is accomplished by the following formulae:

$$w_x(Occ(t_i, q)) = o_i \qquad o_i \in \{0, 1\}. \tag{25}$$

The probability of relevance of document d_j to query q may be then calculated as the value of the numerical term $Rel^{\#}_{d_j}(q)$:

$$Rel^{\#}_{d_j}(q) = \sum_{i=1}^{n} w_x(Occ(t_i, q)) \cdot P^{\#}_{d_j}(t_i). \tag{26}$$

4.2 Representing General Imaging

The $\mathcal{L}_1$ implementation of general imaging differs only slightly from that of standard imaging described in Section 4.1, to reflect the fact that a given keyword may have not one but many equally most similar keywords. This means that, if "most-similarity" values are input from outside, there may be more than one instance of formula (22) for the same t_i. If, instead, "most-similarity" values are to be computed internally, formulae (23) must be substituted by formulae

$$MostSim(t_i, t_{k_i}) \Leftrightarrow \neg\exists t_j.[Sim(t_i, t_j) > Sim(t_i, t_{k_i})]. \tag{27}$$

Each keyword not occurring in the document will now transfer its probability not to a single keyword but to s most similar keywords, in equal parts. The number of keywords that receive some of keyword t_i's probability is expressed by the numeric term

$$\sum_{j=1}^{n} w_x(MostSim(t_i, t_j)) \tag{28}$$

where again we use the observation made in Footnote 6. Numeric terms (24) are then to be substituted by

$$\begin{aligned} w_x(Occ(t_i, d_j)) \cdot \Bigg[& P(t_i) + \\ & \sum_{k=1}^{n} \frac{P(t_k) \cdot w_x(\neg Occ(t_k, d_j)) \cdot w_x(MostSim(t_k, t_i)}{\sum_{l=1}^{n} w_x(MostSim(t_j, t_l))} \Bigg]. \end{aligned} \tag{29}$$

Using the expression $P^{g\#}_{d_j}(t_i)$ as a shorthand of expression (29), the degree of relevance of document d_j to query q may be then calculated as the value of the object term $Rel^{g\#}_{d_j}(q)$:

$$Rel^{g\#}_{d_j}(q) = \sum_{i=1}^{n} w_x(Occ(t_i, q)) \cdot P^{g\#}_{d_j}(t_i). \tag{30}$$

4.3 Representing Proportional Imaging

Also the $\mathcal{L}_1$ implementation of proportional imaging differs only slightly from that described in Section 4.1. What we need to do this time is to reflect the fact that a given keyword not occurring in the document will now transfer its probability neither to a single keyword nor to a subset of most similar keywords, but to all keywords occurring in the document, where the amount of transferred probability is proportional to the degree of similarity between donor and recipient.

In this case formulae (22) and/or (23) are obviously not present. Instead, numerical terms (24) are to be substituted by

$$w_x(Occ(t_i,d_j)) \cdot [P(t_i) + \\ + \sum_{k=1}^{n} P(t_k) \cdot \ w_x(\neg Occ(t_k,d_j)) \cdot \frac{Sim(t_k,t_i)}{\sum_{l=1}^{n}[Sim(t_k,t_l) \cdot w_x(Occ(t_l,d_j))]}]. \tag{31}$$

Using the expression $P^{p\#}_{d_j}(t_i)$ as a shorthand of expression (31), the degree of relevance of document d_j to query q may be then calculated as the value of the numerical term $Rel^{p\#}_{d_j}(q)$:

$$Rel^{p\#}_{d_j}(q) = \sum_{i=1}^{n} w_x(Occ(t_i,q)) \cdot P^{p\#}_{d_j}(t_i). \tag{32}$$

5 RELATED WORK AND DISCUSSION

In this work we have discussed an implementation of a family of non-Bayesian revision methods on top of $\mathcal{L}_1$, a (Bayesian) first order logic extended with features for reasoning about objective probability. This implementation has been achieved by representing the semantic structure that underlies imaging-based conditional logics in the language of $\mathcal{L}_1$. Besides showing the potential of this "Bayesian" tool for the representation of non-Bayesian revision functions, recasting the imaging-related models of information retrieval in such a general purpose knowledge representation and reasoning tool paves the way to a possible integration of these models with other, more KR-oriented models of IR, and to the exploitation of general-purpose domain-knowledge.

The nature of this work may be discussed more effectively by comparing it with the implementation of the imaging-based models of IR discussed in [1, 14]. These works, instead of a full-blown probabilistic FOL, use Probabilistic Datalog [6], an extension of Stratified Datalog (itself a version of the well-known deductive database language Datalog [18]) by means of features for subjective probability. Both in our work and in [1, 14], the entities that participate in the imaging process

(the keywords, their prior probabilities, the similarity values between them, the documents and the queries) are given an explicit representation in the language. Unlike in [1, 14], however, in our approach an explicit representation is given also to the formula that computes the prior probabilities of keywords, to the formula that computes the similarities between keywords and to the formula that chooses the recipients of a probability transfer and computes the revised probabilities of these recipients; the meaning of all these formulae is definable in terms of just the available keyword occurrence data. This hints to the fact that different formulae encoding different methods of computation of the above features may be experimented with in our approach. In this sense, the whole imaging DRF is completely modelled as *a proper theory* of $\mathcal{L}_1$. Instead, the definitions of [14] and [1] are instead rather partial, as most of the reasoning needed for the implementation of the DRF has to be done by some external process.

The approach we propose has the advantage of being more self-contained and conceptually attractive, as it requires the minimum amount of data to be provided from outside the reasoning mechanism. Moreover, with a minimal coding effort, different probability kinematics methods may be experimented with and compared, as can be seen by the ease with which we have encoded the probability transfer formula of different variants of imaging in $\mathcal{L}_1$. The price to be paid for this is that of efficiency, as reasoning in Probabilistic Datalog, a less expressive reasoning tool than $\mathcal{L}_1$, is no doubt more computationally tractable. One may wonder why the implementations of [1, 14] require the prior probabilities of keywords, the similarities between keywords and the revised probabilities of keywords to be computed externally. We think that the answer does not lie in the fact that Probabilistic Datalog is a less powerful tool than $\mathcal{L}_1$, but in the fact that it is inherently geared towards subjective, and not objective, probability (this character of Probabilistic Datalog can be seen from the fact that its semantics contemplates density functions on possible worlds rather than on the individuals of the domain). This entails the impossibility to represent entities that are inherently of a frequentistic nature, such as the IDF of a keyword (see Equation 16) and the notion of similarity between two keywords as degree of coextensionality (see Equation 19). It also somewhat entails a distortion of the meaning of probabilities. For instance, in Probabilistic Datalog one needs to code keyword prior probabilities by means of sentences of type `0.2 term`(t_1), which literally means "the agent believes, with degree of confidence `0.2`, that t_1 is a keyword". In $\mathcal{L}_1$ one writes instead $w_x(t_1) = 0.2$, which means "the probability that a random pick among keywords yields t_1 is 0.2". The latter is no doubt a more faithful rendition of prior probabilities of keywords.

Acknowledgements. I would like to thank Fabio Crestani, Thomas Rölleke and Keith van Rijsbergen for fruitful discussions on the topics of this paper.

REFERENCES

[1] Crestani, F.—Rölleke, Th.: Issues in the implementation of general imaging on top of Probabilistic Datalog. In: M. Lalmas (Ed.): Proceedings of the 1st International Workshop on Logic and Uncertainty in Information Retrieval, Glasgow, UK 1995.

[2] Crestani, F.—Sebastiani, F.—van Rijsbergen, C. J.: Imaging and information retrieval: Variations on a theme. In: F. Crestani, M. Lalmas (Ed.): Proceedings of the 2nd International Workshop on Logic and Uncertainty in Information Retrieval, pp. 48–49, Glasgow, UK 1996.

[3] Crestani, F.—Sebastiani, F.—van Rijsbergen, C. J.: A method for non-binary relevance feedback based on the imaging principle. Technical report, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy 1997 (forthcoming).

[4] Crestani, F.—van Rijsbergen, C. J.: Information retrieval by logical imaging. Journal of Documentation, Vol. 51, 1995, pp. 3–17

[5] Crestani, F.—van Rijsbergen, C. J.: Probability kinematics in information retrieval. In: Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 291–299, Seattle, WA 1995.

[6] Fuhr, N.: Probabilistic Datalog: a logic for powerful retrieval methods. In: Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 282–290, Seattle, WA 1995.

[7] Gärdenfors, P.: Imaging and conditionalization. Journal of Philosophy, Vol. 79, 1982, pp. 747–760

[8] Gärdenfors, P.: Knowledge in flux. Modeling the dynamics of epistemic states. The MIT Press, Cambridge, MA 1988.

[9] Halpern, J. Y.: An analysis of first-order logics of probability. Artificial Intelligence, Vol. 46, 1990, pp. 311–350

[10] Harper, W. L.—Stalnaker, R.—Pearce, G. (Eds.): Ifs. Conditionals, belief, decision, chance and time. Reidel, Dordrecht, NL 1981.

[11] Jeffrey, R. C.: The logic of decision. McGraw Hill, New York, NY 1965.

[12] Lewis, D. K.: Probabilities of conditionals and conditional probabilities. The Philosophical Review, Vol. 85, 1976, pp. 297–315 [a] Also reprinted in [10], pp. 129–147.

[13] Meghini, C.—Sebastiani, F.—Straccia, U.—Thanos, C.: A model of information retrieval based on a terminological logic. In: Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval, pp. 298–307, Pittsburgh, PA 1993. Published by ACM Press, Baltimore, MD.

[14] Rölleke, Th.: Does Probabilistic Datalog meet the requirements of imaging? In: Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, p. 374, Seattle, WA 1995.

[15] Salton, G.: Automatic text processing: the transformation, analysis and retrieval of information by computer. Addison-Wesley, Reading, MA 1989.

[16] Sebastiani, F.: A probabilistic terminological logic for modelling information retrieval. In: Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, pp. 122–130, Dublin, IRL 1994. Published by Springer Verlag, Heidelberg, FRG.

[17] Sebastiani, F.: On the role of logic in information retrieval. Information Processing and Management, 1997 (forthcoming).

[18] Ullman, J. D.: Principles of Database and Knowledge Base Systems, Volume I. Computer Science Press, Potomac, MD 1988.

[19] VAN BENTHEM, J. F.: Exploring logical dynamics. Center for the Study of Language and Information, Stanford, CA 1996.

[20] VAN RIJSBERGEN, C. J.: A non-classical logic for information retrieval. The Computer Journal, Vol. 29, 1986, pp. 481–485

[21] VAN RIJSBERGEN, C. J.: Towards an information logic. In: Proceedings of SIGIR-89, 12th ACM International Conference on Research and Development in Information Retrieval, pp. 77–86, Cambridge, MA 1989.

Manuscript received 8 August 1997

Fabrizio SEBASTIANI (born 1960) graduated in computer science from the University of Pisa, Italy, in 1986, with a thesis dealing with the application of knowledge representation techniques to natural language understanding in technical domains. From 1986 to 1988 he has been a research fellow at the Department of Linguistics, University of Pisa; since 1988 he holds a permanent position as a researcher at IEI-CNR, the Institute for Information Processing of the Italian National Council of Research, Pisa. He has been a visiting research fellow at the Department of Computer Science, University of Toronto, Canada (1989–90), and at the Department of Computer Science, University of Glasgow, UK (1993–94).

He has been active in the area of the logical foundations of artificial intelligence, non-monotonic logics, terminological logics and formalisms for the representation of uncertainty. His most recent interests are in the logical and probabilistic foundations of multimedia information access, including retrieval, filtering and gathering; he has pursued these interests within a number of CEC-funded projects, including the FERMI project ("Formalisation and Experimentation in the Retrieval of Multimedia Information"). He has been lecturing in databases and information systems, and in logical foundations of artificial intelligence, at the universities of Perugia and Pisa, Italy.